Johns Hopkins University, Dept. of Biostatistics Working Papers

6-12-2008

Causal Inference in Observational Studies with Outcome-Dependent Sampling

Weiwei Wang
Johns Hopkins University, wewang@jhsph.edu

Daniel Scharfstein

Zhiqiang Tan

Ellen J. MacKenzie

Causal Inference in Observational Studies with Outcome-Dependent Sampling

Weiwei Wang

The Johns Hopkins Bloomberg School of Public Health, Baltimore, U.S.A.

Daniel Scharfstein

The Johns Hopkins Bloomberg School of Public Health, Baltimore, U.S.A.

Zhiqiang Tan

Rutgers University, Piscataway, U.S.A.

Ellen J. MacKenzie

The Johns Hopkins Bloomberg School of Public Health, Baltimore, U.S.A.

Summary. In this paper, we consider estimation of the causal effect of a treatment on an outcome from observational data collected in two phases. In the first phase, a simple random sample of individuals are drawn from a population. On these individuals, information is obtained on treatment, outcome, and a few low-dimensional confounders. These individuals are then stratified according to these factors. In the second phase, a random sub-sample of individuals are drawn from each stratum, with known, stratum-specific selection probabilities. On these individuals, a rich set of confounding factors are collected. In this setting, we introduce four estimators: (1) simple inverse weighted, (2) locally efficient, (3) doubly robust and (4) enriched inverse weighted. We evaluate the finite-sample performance of these estimators in a simulation study. We also use our methodology to estimate the causal effect of trauma care on in-hospital mortality using data from the National Study of Cost and Outcomes of Trauma.

Keywords: Two-phase sampling; Outcome-dependent sampling; Doubly robust

1. Introduction

In some observational studies, it may be relatively inexpensive to measure the outcome Y, the treatment T and a low-dimensional set of confounders S, while the remaining set of covariates W are expensive to obtain. In such settings, an outcome-dependent two-phase sampling design, first introduced by Neyman (1938) and discussed in Cochran (1963), can significantly reduce the cost of the study. In the first phase, Y, T and S are collected on all subjects. In the second phase, the subjects are divided into strata according to (Y, T, S) and a random subsample, also called a validation sample, is drawn from each strata. The expensive covariates W are measured only on the validation sample.

Existing statistical methods for outcome-dependent sampling studies are primarily focused on obtaining consistent and efficient estimates for the regression parameters in a population-level model for the distribution for Y given T, S and W, see, for example, Cosslett (1981, 1983), White (1982), Fears and Brown (1986), Breslow and Cain (1988), Pepe and Fleming (1991), Carroll and Wand (1991), Schill et al. (1993), Scott and Wild (1991, 1997), Robins et al. (1995), Breslow and Holubkov (1997a,b), Lawless et al. (1999), Breslow (2000), Breslow et al. (2003), Chatterjee et al. (2003), Weaver and Zhou (2005).

In constrast to the above methods, we are interested in estimating the causal effect of a non-randomized treatment on the outcome. Specifically, we want to obtain estimators for the marginal mean of the "potential" outcome that would have been observed had all subjects received a specified treatment. Our work is motivated, in part, by the National Study on the Costs and Outcomes of Trauma (NSCOT; MacKenzie *et al.*, 2006), in which an outcome-dependent sample of severely injured patients treated at trauma and non-trauma centeres are prospectively followed for survival and functional status. One of the primary aims is to estimate the causal effect of trauma center vs. non-trauma center care on in-hospital mortality, accounting for (1) outcome dependent sampling and (2) selection bias due to non-randomized assignment of type of care.

The paper is organized as follows. Section 2 introduces the notation and data structure. Section 3 introduces four estimators: (1) simple inverse weighted (SIW), (2) locally efficient (LE), (3) doubly robust (DR) and (4) enriched inverse weighted (EIW) which is more efficient than the SIW estimator. Section 4 provides the results of a comprehensive simulation study. Section 5 presents an analysis of the NSCOT data using our estimators. The last section is devoted to a discussion.

2. Notation and Framework

For an individual, let $X = (S', W')'$ denote covariates and let Y be the observed outcome. On all subjects, $(S', Y, T)'$ is collected while W is only collected on the validation sample. Let V be the validation indicator and T be the treatment indicator. Based on the study design, the probability of being part of the validation sample depends on (S, Y, T), but not W. We denote $q(S, Y, T) = P(V = 1|S, Y, T)$, which is known by study design. For clarity, we further denote $q_1(S, Y) = q(S, Y, 1)$ and $q_0(S, Y) = q(S, Y, 0)$. Let Y_1 and Y_0 be the potential outcomes under treatment 1 and 0, respectively. We make the stable unit treatment assumption (Rubin, 1986), so that $Y = TY_1 + (1 - T)Y_0$.

The observed data for an individual is $O = (S', Y, T, V, V \cdot W')'$. We are interested in using n i.i.d. copies of O to draw inference about $\mu_t^* = E[Y_t]$. To identify μ_t^* from the observed data, we assume that there are *no unmeasured confounders*, i.e., T is independent of Y_t given X. We let $\pi_t^*(X) = P[T = t|X]$. We assume that $\pi_t^*(x) > 0$ for all x and $t = 0, 1$. For convenience, Table 1 summarizes our notation. When it is important, we will emphasize the dependence of functions on model parameters; otherwise, we will suppress such dependence.

3. Estimation

3.1. Simple Inverse Weighted Estimation

Under the above assumptions, it can be shown that the following observed data estimating function:

$$U_t^{SIW}(O; \mu_t, \pi_t) = \frac{VI_{\{T=t\}}}{q_t(S, Y)\pi_t(X)}(Y - \mu_t)$$

has mean 0 when evaluated at μ_t^* and $\pi_t^*(X)$. Without additional modeling assumptions, this estimating function cannot be used to draw $\sqrt{n}$ inference about μ_t^*. This is because one would need a non-parametric estimator of $\pi_t^*(X)$ converging at a rate faster than $n^{1/4}$, which cannot be achieved when X is high-dimensional (Newey, 1994). This has been called the "curse of dimensionality" (Robins and Ritov, 1997).

Table 1. Notation

Symbol	Description
Y	Observed outcome
T	Treatment (0/1)
Y_1	Potential outcome under treatment 1
Y_0	Potential outcome under treatment 0
μ_1^*	$E[Y_1]$
μ_0^*	$E[Y_0]$
S	Covariates observed for all subjects
W	Covariates observed only on validation sample
X	$(S', W')'$
V	Validation indicator
$\mu_1^*(X)$	$E[Y_1\|X]$
$\mu_0^*(X)$	$E[Y_0\|X]$
$q(S,Y,T)$	$Pr(V=1\|S,Y,T)$
$q_1(S,Y)$	$q(S,Y,1)$
$q_0(S,Y)$	$q(S,Y,0)$
$\pi_1^*(X)$	$Pr[T=1\|X]$
$\pi_0^*(X)$	$Pr[T=0\|X]$

In order to proceed, one can parameterically model $\pi_t^*(X) = \pi_t(X;\alpha^*)$, where α^* denotes the true value of the model parameters α. Without loss of generality, we will assume that

$$\text{logit } \pi_1(X;\alpha^*) = l(X;\alpha^*)$$

where $l(X,\alpha)$ is a specified function of X and α, which is differentiable in α. This model implies that $\pi_t(X;\alpha^*) = \exp(t\cdot l(X;\alpha^*))/(1+\exp(l(X;\alpha^*)))$. Under this model, α^* can be estimated as the solution, $\widehat{\alpha}$, to the logistic regression T given X score function where each individual gets weight $\frac{V}{q_T(S,Y)}$. Let

$$S_\alpha(T,X;\alpha) = \frac{\partial l(X;\alpha)}{\partial \alpha}(T-\pi_1(X;\alpha)),$$

then $\hat{\alpha}$ is the solution to

$$E_n\left[\frac{V}{q_T(S,Y)}S_\alpha(T,X;\alpha)\right] = 0,$$

where $E_n[\cdot]$ is the empirical average operator.

In terms of estimating μ_t^*, we note that

$$U_t^{SIW}(O;\mu_t,\alpha) = \frac{VI_{\{T=t\}}}{q_t(S,Y)\pi_t(X;\alpha)}(Y-\mu_t)$$

has mean zero when evaluated at μ_t^* and α^*. This suggests that one can estimate μ_t^* as the solution $E_n[U_t^{SIW}(O;\mu_t,\widehat{\alpha})] = 0$. The resulting estimator is then

$$\widehat{\mu}_t^{SIW} = E_n\left[\frac{VI_{\{T=t\}}}{q_t(S,Y)\pi_t(X;\widehat{\alpha})}Y\right] / E_n\left[\frac{VI_{\{T=t\}}}{q_t(S,Y)\pi_t(X;\widehat{\alpha})}\right]$$

The influence function for the inverse weighted estimator is

$$\begin{aligned} IF_t^{SIW}(O;\mu_t^*,\alpha^*) &= U_t^{SIW}(O;\mu_t^*,\alpha^*) - \\ & E\left[\frac{\partial U_t^{SIW}(O;\mu_t,\alpha^*)}{\partial\alpha}\right] E\left[\frac{V}{q_T(S,Y)}\frac{\partial S_\alpha(T,X;\alpha^*)}{\partial\alpha}\right]^{-1} \\ & \times\frac{V}{q_T(S,Y)}S_\alpha(T,X;\alpha^*). \end{aligned}$$

The asymptotic variance of the inverse weighted estimator can be estimated by $E_n[\widehat{IF}_t^{SIW}(O;\widehat{\mu}_t^{SIW},\widehat{\alpha})^2]$, where $\widehat{IF}_t^{SIW}(O;\mu_t,\alpha)$ is the same as $IF_t^{SIW}(O;\mu_t,\alpha)$ except that the expectations are replaced with empirical averages.

The advantage of this simple inverse weighted estimator is that it is simple to compute and only requires modeling of $\pi_t^*(X)$. The disadvantage of this estimator is two-fold. First, it does not take full advantage of the data recorded on subjects who are not in the validation sample. It only uses information on (S,Y,T) through the construction of the sampling weights. Second, the simple inverse weighted estimator is inefficient, even if we ignore data on those who are not validated.

3.2. Representation of Influence Functions

Using the semiparametric theory of inference in coarsened data problems (Bickel et al., 1993; Robins et al., 1994; Scharfstein et al., 1999; van der Laan and Robins, 2003; Tsiatis, 2006), we derive the class of all influence functions for regular and asymptotically linear (RAL) estimators of μ_t^* under correct specification of the logistic regression model for $\pi_t^*(X)$ discussed in the previous subsection. In the Appendix, we show that the class of all influence functions consists of elements of the form:

$$IF_t(O;\mu_t^*,\alpha^*,\mu_t^*(X),g_1,g_0,b) = \phi_1(O;\mu_t^*,\alpha^*,\mu_t^*(X)) + \sum_{\tau=0}^{1}\phi_{2,\tau}(O;g_\tau) + \phi_3(O;\alpha^*,b) \quad (1)$$

where

$$\begin{aligned} \phi_1(O;\mu_t,\alpha,\mu_t(X)) &= \frac{VI_{\{T=t\}}}{q_t(S,Y_t)\pi_t(X;\alpha)}(Y_t-\mu_t) + \\ & (-1)^t\frac{V}{q_T(S,Y_T)}\left(\frac{T-\pi_1(X;\alpha)}{\pi_t(X;\alpha)}\right)(\mu_t(X)-\mu_t) \\ \phi_{2,\tau}(O;g_\tau) &= I_{\{T=\tau\}}\left(1-\frac{V}{q_\tau(S,Y_\tau)}\right)g_\tau(S,Y_\tau) \\ \phi_3(O;\alpha^*,b) &= \frac{V}{q_T(S,Y_T)}\frac{T-\pi_1(X;\alpha^*)}{R(X;\alpha^*)}b(X), \end{aligned}$$

g_1, g_0 are arbitrary functions of their arguments, $R(X;\alpha^*) = \pi_1(X;\alpha^*)\pi_0(X;\alpha^*)$, $b(X) \in \mathcal{T}^\perp$ and

$$\mathcal{T} = \left\{k'\frac{\partial l(X;\alpha^*)}{\partial\alpha} : k \text{ is an arbitrary constant vector}\right\}.$$

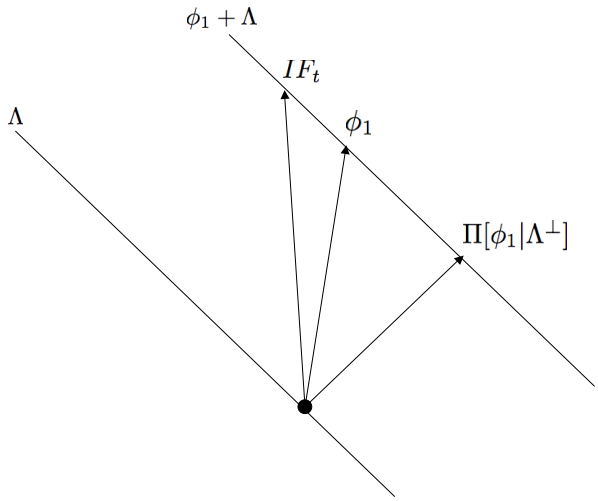

Fig. 1. Optimal influence function

The influence function for the inverse weighted estimator is a member of this class with $g_1(S, Y_1) = 0$ and $g_0(S, Y_0) = 0$ and $b(X)$ equal to

$$\begin{aligned} b^{SIW}(X) &= (-1)^{t+1}\pi_{1-t}(X;\alpha^*)(\mu_t^*(X) - \mu_t^*) - \\ &\quad R(X;\alpha^*)E\left[\pi_{1-t}(X;\alpha^*)(Y_t - \mu_t^*)\frac{\partial l(X;\alpha^*)}{\partial \alpha}\right] \times \\ &\quad E\left[R(X,\alpha^*)\frac{\partial l(X;\alpha^*)}{\partial \alpha}\frac{\partial l(X;\alpha^*)}{\partial \alpha}'\right]^{-1}\frac{\partial l(X;\alpha^*)}{\partial \alpha}. \end{aligned}$$

Letting

$$\begin{aligned} \Lambda_{g_\tau} &= \{\phi_{2,\tau}(O;g_\tau) : g_\tau \text{ is arbitrary}\} \\ \Lambda_b &= \left\{\phi_3(O;\alpha^*,b) : b(X) \in \mathcal{T}^\perp\right\}. \end{aligned}$$

$\Lambda_g = \Lambda_{g_0} \oplus \Lambda_{g_1}$, and $\Lambda = \Lambda_g + \Lambda_b$, the class of all influence functions can be written as a linear variety of form:

$$\{IF_t(O;\mu_t^*,\alpha^*,\mu_t^*(X),g_1,g_0,b) : g_1,g_0,b\} = \phi_1(O;\mu_t^*,\alpha^*,\mu_t^*(X)) + \Lambda$$

Note that Λ_{g_0} and Λ_{g_1} are orthogonal, while Λ_g and Λ_b are not.

3.3. Locally Efficient Estimation

The variance of an influence function in the above class depends on the choice of g_1, g_0, and $b(X)$. The influence function with the smallest variance (i.e., most efficient) is the projection $\Pi[\phi_1(O;\mu_t^*,\alpha^*,\mu_t^*(X))|\Lambda^\perp]$, where $\Pi[\cdot|\cdot]$ is the projection operator (Figure 1).

The required projection does not have a closed form solution, however one can use the alternating projection theorem (Bickel et al., 1993) to computationally approximate the solution. We reproduce a version of the theorem here.

Alternating Projection Theorem: *Let S_A and S_B be two non-orthogonal spaces. Let h denote a vector. Let $Q[\cdot]$, $Q_A[\cdot]$ and $Q_B[\cdot]$ be the projection operator to the spaces $(S_A + S_B)^\perp$, $S_A^\perp$ and $S_B^\perp$, respectively. Let $Q^{(1)}[h] = Q_B[Q_A[h]]$ and $Q^{(j)}[h] = Q^{(1)}[Q^{(j-1)}[h]]$, $j > 1$. Then,*

$$\lim_{j\to\infty} Q^{(j)}[h] = Q[h].$$

In our context $S_A = \Lambda_g$, $S_B = \Lambda_b$, and $h(O) = \phi_1(O; \mu_t^*, \alpha^*, \mu_t^*(X))$. Figure 2 illustrates the alternating projection theorem in our context. The key then is to derive closed form expressions for the projection operators Q_A and Q_B. To do so, we need to know how to project a function of the observed data $h(O)$ onto Λ_g and onto Λ_b. It can be shown that

$$\Pi[h(O)|\Lambda_g] = \sum_{\tau=0}^{1} I_{\{T=\tau\}} \left(1 - \frac{V}{q_\tau(S, Y_\tau)}\right) \frac{E\left[h(O) I_{\{T=\tau\}} \left(1 - \frac{V}{q_\tau(S,Y_\tau)}\right) \,\middle|\, S, Y_\tau\right]}{E\left[I_{\{T=\tau\}} \left(1 - \frac{V}{q_\tau(S,Y_\tau)}\right)^2 \,\middle|\, S, Y_\tau\right]} \tag{2}$$

and

$$\Pi[h(O)|\Lambda_b] = \frac{V}{q_T(S, Y_T)} \left(\frac{T - \pi_1(X; \alpha^*)}{R(X; \alpha^*)}\right) A(X)^{-1} \left\{C_h(X) - k_h' \frac{\partial l(X; \alpha^*)}{\partial \alpha}\right\}, \tag{3}$$

where

$$C_h(X) = E\left[h(O) \frac{V}{q_T(S, Y_T)} \frac{T - \pi_1(X; \alpha^*)}{R(X; \alpha^*)} \,\middle|\, X\right],$$

$$A(X) = \pi_1(X; \alpha^*)^{-1} E\left[q_1(S, Y_1)^{-1} \,|\, X\right] + \pi(0, X; \alpha^*)^{-1} E\left[q_0(S, Y_0)^{-1} \,|\, X\right]$$

and

$$k_h' = E\left[A(X)^{-1} C_h(X) \frac{\partial l(X; \alpha^*)}{\partial \alpha}'\right] E\left[A(X)^{-1} \frac{\partial l(X; \alpha^*)}{\partial \alpha} \frac{\partial l(X; \alpha^*)}{\partial \alpha}'\right]^{-1}.$$

To find the projection, $\Pi[\phi_1(O; \mu_t^*, \alpha^*, \mu_t^*(X))|\Lambda^\perp]$, we use the following algorithm. Let $h_t^{(j)}(O)$ denote the result of the jth iteration.

- Compute

$$h_t^{(1)}(O) = \phi_1(O; \mu_t^*, \alpha^*, \mu_t^*(X)) - \Pi[\phi_1(O; \mu_t^*, \alpha^*, \mu_t^*(X))|\Lambda_g] \tag{4}$$

- Compute

$$\begin{aligned} h_t^{(2)}(O) =& \phi_1(O; \mu_t^*, \alpha^*, \mu_t^*(X)) - \Pi[\phi_1(O; \mu_t^*, \alpha^*, \mu_t^*(X))|\Lambda_g] \\ & - \Pi[\phi_1(O; \mu_t^*, \alpha^*, \mu_t^*(X))|\Lambda_b] + \Pi[\Pi[\phi_1(O; \mu_t^*, \alpha^*, \mu_t^*(X))|\Lambda_g]|\Lambda_b] \end{aligned} \tag{5}$$

- On iteration $2k + 1$ $(k > 1)$, compute

$$h_t^{(2k+1)}(O) = h_t^{(2k)}(O) + \Pi[\Pi[h_t^{(2k-1)}(O)|\Lambda_b]|\Lambda_g]$$

- On iteration $2k + 2$ $(k > 1)$, compute

$$h_t^{(2k+2)}(O) = h_t^{(2k+1)}(O) + \Pi[\Pi[h_t^{(2k)}(O)|\Lambda_g]|\Lambda_b]$$

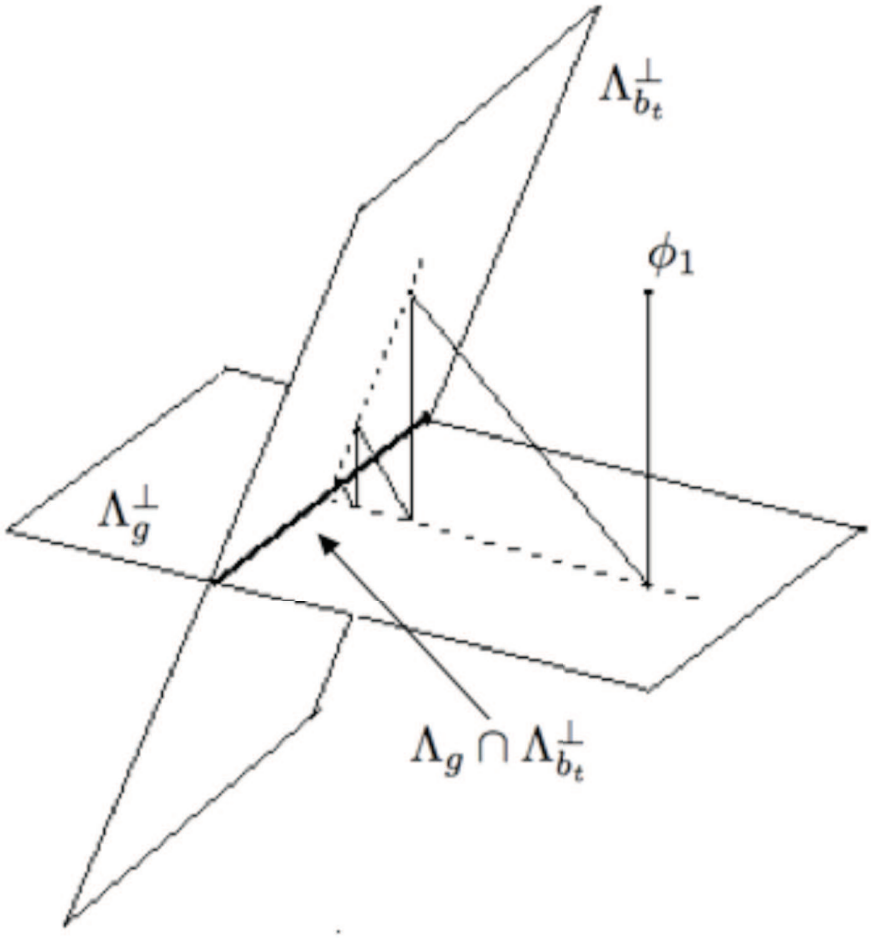

Fig. 2. Alternating Projection Theorem

While the general projection formulae are given above, the following special cases are used in the algorithm. When $h(O) = \phi_1(O; \mu_t^*, \alpha^*, \mu_t^*(X))$ (as in iterations 1 and 2), then $\Pi[\phi_1(O; \mu_t^*, \alpha^*, \mu_t^*(X))|\Lambda_g]$ takes the form in Equation (2) where the ratio of the conditional expectations is equal to

$$-\frac{I_{\{\tau=t\}}(Y_t - \mu_t^*) + (-1)^t E\left[\frac{\pi_\tau(X;\alpha^*)}{\pi_t(X;\alpha^*)}(\tau - \pi_1(X;\alpha^*))(\mu_t^*(X) - \mu_t^*)\,\middle|\, S, Y_\tau\right]}{E[\pi_\tau(, X; \alpha^*)|S, Y_\tau]}$$

and $\Pi[\phi_1(O; \mu_t^*, \alpha^*, \mu_t^*(X))|\Lambda_b]$ takes the form in Equation (3) with

$$C_{\phi_1}(X) = \frac{(-1)^{t+1}}{\pi_t(X;\alpha^*)} E\left[\frac{Y_t - \mu_t^*}{q_t(S, Y_t)}\,\middle|\, X\right] + (-1)^t A(X)\pi(1-t;\alpha^*)(\mu_t^*(X) - \mu_t^*).$$

When $h(O) = \phi_3(O; \alpha^*, b^\dagger) \in \Lambda_b$ (as in the odd iterations greater than 1), then $\Pi[\phi_3(O; \alpha^*, b^\dagger)|\Lambda_g]$ takes the form in Equation (2) with the ratio of the conditional expectations for τ equal to

$$(-1)^\tau E[\pi_\tau(X;\alpha^*)|S, Y_\tau]^{-1} E[b^\dagger(X)\,|\,S, Y_\tau].$$

When $h(O) = \sum_{\tau=0}^1 \phi_{2,\tau}(O; g_\tau^\dagger) \in \Lambda_g$ (as in all even iterations), then $\Pi\left[\sum_{\tau=0}^1 \phi_{2,\tau}(O; g_\tau^\dagger)\,\middle|\,\Lambda_b\right]$ takes the form in Equation (3) with

$$C_{\sum_{\tau=0}^1 \phi_{2,\tau}}(X) = \sum_{\tau=0}^{1} (-1)^{\tau+1} E\left[g_\tau^\dagger(S, Y_\tau)\left(1 - \frac{1}{q_\tau(S, Y_\tau)}\right)\,\middle|\, X\right].$$

On each iteration, $h_t^{(j)}(O)$ can be expressed as $IF_t(O; \mu_t^*, \alpha^*, g_1^{(j)}, g_0^{(j)}, b^{(j)})$ for $g_1^{(j)}, g_0^{(j)}, b^{(j)}$ determined via the above projection formulae. Thus, for j^* large, the solution to

$$E_n[IF_t(O; \mu_t, \widehat{\alpha}, \mu_t^*(X), g_1^{(j^*)}, g_0^{(j^*)}, b^{(j^*)})] = 0$$

will be an efficient estimator of μ_t^* whose influence function approximates $\Pi[\phi_1(O; \mu_t^*, \alpha^*, \mu_t^*(X))|\Lambda^\perp]$. This estimator, however, depends on $\mu_t^*(X)$ and $(g_1^{(j)}, g_0^{(j)}, b^{(j)})$, which as seen in the preceeding displays, depends on conditional expectations of the form $E[u(X, Y_t)|S, Y_t]$ and $E[u(X, Y_t)|X]$ and marginal expectations of the form $E[u(X, Y_t)]$ and $E[v(X)]$ where $u(\cdot)$ and $v(\cdot)$ are functions of (X, Y_t) and X respectively.

To proceed, we will need to utilize a *working* model for $\mu_t^*(X)$. Regardless of whether this working model is correctly specified and how many iterations are used to compute the projection, the resulting estimators will be consistent, since it can be shown that $E[IF_t(O; \mu_t^*, \alpha^*, f(X), g_1, g_0, b)] = 0$, whatever be the choice of f, g_1, g_0, and b. However, if this working model is correctly specified, its parameters can be estimated at $n^{1/4+\epsilon}$ rates, and the number of iterations in the projection are large, the resulting estimator will be efficient, achieving the semiparametric variance bound. As a result, our estimator is said to be "locally" efficient.

3.3.1. Binary Outcomes

For a binary outcome, we can posit a working logistic regression model for $\mu_t^*(X) = \mu_t(X; \eta^*)$, where

$$\text{logit } \mu_t(X; \eta^*) = j(t, X; \eta^*)$$

and $j(t, X; \eta)$ is a specified function of t, X, and η. To estimate η^*, we notice that

$$\mu_t(X; \eta^*) = E[Y_t|X, T = t] = E[Y|X, T = t],$$

where the first equality follows by the assumption of no unmeasured confouders and the second by the stable unit treatment assumptoin. Thus, η^*, can be estimated as the solution, $\widehat{\eta}$, to the re-weighted Y_t given X score equations where each individual receives weight $\frac{VI_{\{T=t\}}}{q_T(S,Y_T)}$. Let

$$S_{t\eta}(Y, X; \eta) = \frac{\partial j(t, X; \eta)}{\partial \eta}(Y - \mu_t(X; \eta)),$$

then $\hat{\eta}$ is the solution to

$$E_n\left[\frac{VI_{\{T=t\}}}{q_T(S, Y_T)} S_{t\eta}(Y, X; \eta)\right] = 0.$$

The projections above require estimation of conditional and marginal expectations of the form: $E[u(X, Y_t)|X]$, $E[u(X, Y_t)|S, Y_t]$, $E[u(X, Y_t)]$ and $E[v(X)]$. We can estimate the conditional expectation $E[u(X, Y_t)|X]$ by $u(X, 1)\mu_t(X; \widehat{\eta}) + u(X, 0)(1 - \mu_t(X; \widehat{\eta}))$, where $\mu_t(X; \eta) = \exp(j(t, X; \eta))/(1 + \exp(j(t, X; \eta)))$. This estimator presumes, as we do, that $u(x, 1)$ and $u(x, 0)$ are well defined for all x. The conditional expectation $E[u(X, Y_t)|S, Y_t]$ can be estimated by weighted, saturated regression with individual weights $\frac{VI_{\{T=t\}}}{q_t(S,Y_t)\pi_t(X;\widehat{\alpha})}$. The marginal expectations $E[u(X, Y_t)]$ and $E[v(X)]$ can be estimated by weighted averages with individual weights $\frac{VI_{\{T=t\}}}{q_t(S,Y_t)\pi_t(X;\widehat{\alpha})}$ and $\frac{V}{q_T(S,Y_T)}$, respectively.

Our resulting estimator will be the solution, $\widehat{\mu}_t^{(j^*)}$, to

$$E_n[IF_t(O; \mu_t, \widehat{\alpha}, \mu_t(X; \widehat{\eta}), \widehat{g}_1^{(j^*)}, \widehat{g}_0^{(j^*)}, \widehat{b}^{(j^*)})] = 0$$

where $\widehat{g}_1^{(j^*)}, \widehat{g}_0^{(j^*)}, \widehat{b}^{(j^*)}$ are obtained by replacing conditional and marginal expectations by their estimates described above, and j^* is an iteration number in which the change in the absolute difference between $\widehat{\mu}_t^{(j^*)}$ and $\widehat{\mu}_t^{(j^*-1)}$ is smaller than a specified tolerance ϵ. We define $\widehat{\mu}_t^{LE} = \widehat{\mu}_t^{(j^*)}$. When the working model for $\mu_t^*(X)$ is correctly specified, the influence function of $\widehat{\mu}_t^{(j^*)}$ will be $IF_t(O; \mu_t^*, \alpha^*, \mu_t(X; \eta^*), g_1^{(j^*)}, g_0^{(j^*)}, b^{(j^*)})$. The asymptotic variance of this estimator can be estimated by $E_n[IF_t(O; \widehat{\mu}_t^{LE}, \widehat{\alpha}, \mu_t(X; \widehat{\eta}), \widehat{g}_1^{(j^*)}, \widehat{g}_0^{(j^*)}, \widehat{b}^{(j^*)})^2]$.

If, however, the working model for $\mu_t^*(X)$ is incorrectly specified, then $\widehat{\eta}$ converges to $\tilde{\eta} \neq \eta^*$. Consequently, $IF_t(O; \mu_t^*, \alpha^*, \mu_t(X; \tilde{\eta}), g_1, g_0, b)$ is not an influence function because it is no longer orthogonal to the nuisance tangent space for the propensity score. In order to get the influence function, we must subtract its projection onto the space spanned by the score functions for the propensity score, which is equal to

$$\begin{aligned} IF_t^R(O; \mu_t^*, \alpha^*, \tilde{\eta}, g_1^{(j^*)}, g_0^{(j^*)}, b^{(j^*)}) =& IF_t(O; \mu_t^*, \alpha^*, \mu_t(X; \tilde{\eta}), g_1^{(j^*)}, g_0^{(j^*)}, b^{(j^*)}) \\ & - E\left[\frac{\partial IF_t(O; \mu_t^*, \alpha^*, \mu_t(X; \tilde{\eta}), g_1^{(j^*)}, g_0^{(j^*)}, b^{(j^*)})}{\partial \alpha}\right] \\ & \times E\left[\frac{V}{q_T(S, Y)} \frac{\partial S_\alpha(T, X; \alpha)}{\partial \alpha}\right]^{-1} \\ & \times \frac{V}{q_T(S, Y)} S_\alpha(T, X; \alpha). \end{aligned} \tag{6}$$

It is important to note that the second term in this influence function will be zero if $\tilde{\eta} = \eta^*$. We estimate the asymptotic variance of $\hat{\mu}^{LE}$ by $E_n[\widehat{IF}_t^R(O; \widehat{\mu}_t^{LE}, \widehat{\alpha}, \widehat{\eta}, \widehat{g}_1^{(j^*)}, \widehat{g}_0^{(j^*)}, \widehat{b}^{(j^*)})^2]$, where $\widehat{IF}_t^R(O; \mu_t^*, \alpha^*, \tilde{\eta}, g_1^{(j^*)}, g_0^{(j^*)}, b^{(j^*)})$ is the same as $IF_t^R(O; \mu_t^*, \alpha^*, \tilde{\eta}, g_1^{(j^*)}, g_0^{(j^*)}, b^{(j^*)})$ except that the expectations are replaced by emprical averages. The asymptotic variance calculated using Equation (6) and its associated estimator will be refered to as the "robust variance" since it is robust to misspecification of the working model for $\mu_t^*(X)$.

3.3.2. Continuous Outcomes

For a continuous outcome, one can specify a working model for $\mu_t^*(X)$ in one of two ways. First, one can specify a fully parametric model for the conditional distribution of Y_t given X and estimate the parameters, η^*, using weighted score equations as above. Second, one can specify a (semi-parametric) conditional mean model for Y_t given X and estimate the parameters, η^*, using weighted estimating equations. In this latter approach, we could define $S_{t\eta}(Y, X; \eta) = \frac{\partial \mu_t(X;\eta)}{\partial \eta}(Y - \mu_t(X; \eta))$.

One must also estimate of conditional expectations of the form $E[u(X, Y_t)|X]$. If the first approach of specifying model for $\mu_t^*(X)$ is adopted, then one can estimate these expectations by using the estimated conditional distribution of Y_t given X. Since the second approach leaves the distribution form of Y_t given X unspecified, one must introduce additional modeling assumptions here. Specifically, one can model the conditional expectations directly by specifying a regression model for $u(X, Y_t)$ given X and estimate the parameters using additional weighted estimating equations. As the conditional expectations of this form must be repeatedly estimated, this approach can lead to issues of model incompatability. The first approach does not suffer from this issue. If, from either approach, the resulting design matrix for the regression is the same as that for the propensity score model, the right hand side of Equation (3) will be equal to zero and the locally efficient estimator will reduce to the doubly robust estimator (see next section).

For a continuous outcome, smoothing will be required to estimate $E[v(X)|S, Y_t]$. If S is relatively low-dimensional, one can perform non-parametric smoothing in Y_t within levels of S. The stratum-specific non-parametric estimators must be $n^{1/4+\epsilon}$ consistent. If S has many levels, one may need to additionally smooth across levels of S and perform parametric or semi-parametric regression. In these procedures, individuals are weighted with individual weights $\frac{VI_{\{T=t\}}}{q_t(S,Y_t)\pi_t(X;\widehat{\alpha})}$.

As in binary outcomes, if the working model for $\mu_t^*(X)$ is misspecified, the robust variance should be calculated using Equation (6). Misspecification of $E[u(X, Y_t)|X]$ or $E[v(X)|S, Y_t]$ will not affect the consistency of the resulting estimator, just the efficiency.

3.4. Doubly Robust Estimation

In this section, we consider the sub-class of influence functions with $b(X) = 0$, i.e. $\phi_1(O; \mu_t^*, \alpha^*, \mu_t^*(X)) + \Lambda_g$. This class is generated by the model in which no restrictions are placed on the distribution of $\pi_t^*(X)$. Interestingly, this class of influence functions has the following properties:

(a) $E[IF_t(O; \mu_t^*, \tilde{\alpha}, \mu_t(X; \eta^*), g_1, g_0, 0)] = 0$ whatever be $\tilde{\alpha}$
(b) $E[IF_t(O; \mu_t^*, \alpha^*, \mu_t(X; \tilde{\eta}), g_1, g_0, 0)] = 0$ whatever be $\tilde{\eta}$

As a result, members of this subclass are said to be "doubly robust" in the sense that the resulting estimators will be consistent and asymptotically normal when either the model for $\pi_t^*(X)$ or the model for $\mu_t^*(X)$ is correctly specified. In our setting, there are an infinite number of doubly robust estimators, depending on the selection of the functions g_1 and g_0. The estimator with the smallest variance will have influence function equal to the projection of $\phi_1(O;\mu_t^*,\alpha^*,\mu_t(X;\eta^*))$ onto $\Lambda_g^{\perp}$. Using previous results, this projection is equal to $h_t^{(1)}(O;\mu_t^*,\alpha^*,\mu_t(X;\eta^*))$ in Equation (4), which is the first iteration in the alternating projection algorithm. The previous subsection has provided a detailed illustration of the computation of $h_t^{(1)}(O;\mu_t^*,\alpha^*,\mu_t(X;\eta^*))$. It only involves conditional expectations of the form $E[v(X)|S,Y_t]$ and is computationally easier than the locally efficient estimator. While mis-specification of $E[v(X)|S,Y_t]$ may result in loss of efficiency, the estimator of μ_t^* will still remain doubly robust.

We denote the most efficient doubly robust estimator as $\widehat{\mu}_t^{DR}$. The locally efficient estimator is not doubly robust, in contrast to many other cases discussed in van der Laan and Robins (2003).

When either the model for $\pi_t^*(X)$ or the model for $\mu_t^*(X)$ is incorrectly specified, the influence function for the resulting estimator can be shown to be

$$\begin{aligned} IF_t^{DR}(O;\mu_t^*,\tilde{\alpha},\tilde{\eta}) =& h_t^{(1)}(O;\mu_t^*,\tilde{\alpha},\mu_t(X;\tilde{\eta})) \\ &- E\left[\frac{\partial h_t^{(1)}(O;\mu_t^*,\tilde{\alpha},\mu_t(X;\tilde{\eta}))}{\partial\alpha}\right] E\left[\frac{V}{q_T(S,Y)}\frac{\partial S_\alpha(T,X;\tilde{\alpha})}{\partial\alpha}\right]^{-1} \\ &\times \frac{V}{q_T(S,Y)} S_\alpha(T,X;\tilde{\alpha}) \\ &- E\left[\frac{\partial h_t^{(1)}(O;\mu_t^*,\tilde{\alpha},\mu_t(X;\tilde{\eta}))}{\partial\eta}\right] E\left[\frac{VI_{\{T=t\}}}{q_T(S,Y)}\frac{\partial S_{t\eta}(Y,X;\tilde{\eta})}{\partial\eta}\right]^{-1} \\ &\times \frac{VI_{\{T=t\}}}{q_T(S,Y)} S_{t\eta}(Y,X;\tilde{\eta})'. \end{aligned} \tag{7}$$

where either $\tilde{\alpha}=\alpha^*$ (propensity model correct) or $\tilde{\eta}=\eta^*$ (outcome regression correct). In Equation (7), the second term vanishes if $\tilde{\eta}=\eta^*$ (outcome regression model correct) and the third term vanishes if $\tilde{\alpha}=\alpha^*$ (propensity model correct). We estimate the asymptotic variance of $\hat{\mu}^{DR}$ by $E_n[\widehat{IF}_t^{DR}(O;\widehat{\mu}_t^{DR},\widehat{\alpha},\widehat{\eta})^2]$, where $\widehat{IF}_t^{DR}(O;\mu_t^*,\tilde{\alpha},\tilde{\eta})$ is the same as $IF_t^{DR}(O;\mu_t^*,\tilde{\alpha},\tilde{\eta})$ except that the expectations are replaced by emprical averages. The asymptotic variance calculated using Equation (7) and its associated estimator will be refered to as the "doubly robust variance" since it is robust to misspecification of the model for $\mu_t^*(X)$ or the model for $\pi_t^*(X)$.

3.5. Enriched Inverse Weighted Estimator

When compared to the simple inverse weighted estimator, the locally efficient estimator will always have a smaller variance but it is computationally intensive. The doubly robust estimator is computationally easier and has the nice property of double robustness, but it is not guaranteed to have a smaller variance than the inverse weighted estimator and a working model for $\mu_t^*(X)$ is required. This motivates our search for a computationally feasible estimator which always has a smaller variance than the simple inverse weighted

estimator. Here, we restrict attention to the influence functions that are elements of the linear variety: $IF_t^{SIW}(O;\mu_t^*,\alpha^*)+\Lambda_g$. Among this set of influence functions, the most efficient one is equal to the projection $IF_t^{SIW}(O;\mu_t^*,\alpha^*)$ onto $\Lambda_g^{\perp}$. We define

$$IF_t^{EIM}(O;\mu_t^*,\alpha^*)=IF_t^{SIW}(O;\mu_t^*,\alpha^*)-\Pi[IF_t^{SIW}(O;\mu_t^*,\alpha^*)|\Lambda_g].$$

Equation (2) shows how to project a general function of observed data onto Λ_g. When $h(O)=\widehat{IF}_t^{SIW}(O;\mu_t^*,\alpha^*)$, the ratio of the conditional expectations for τ in Equation (2) takes the form

$$-\frac{I_{\{\tau=t\}}(Y_t-\mu_t^*)}{E[\pi(\tau,X;\alpha^*)|S,Y_\tau]} - \frac{(-1)^\tau E\left[\frac{\partial U_t^{SIW}(O;\mu_t^*,\alpha^*)}{\partial\alpha}\right]E\left[\frac{V}{q_T(S,Y)}\frac{\partial S_\alpha(T,X;\alpha^*)}{\partial\alpha}\right]^{-1}E\left[\frac{\partial l(X;\alpha^*)}{\partial\alpha}R(X;\alpha)|S,Y_\tau\right]}{E[\pi(\tau,X;\alpha^*)|S,Y_\tau]}$$

The enriched inverse weighted estimator does not require a model for $\mu_t^*(X)$. It does, however, require estimation of conditional expectations of the form $E[v(X)|S,Y_t]$ (see Sections 3.3.1 and 3.3.2 for estimation strategies; model mis-specification may result in loss of efficiency, in which case the enriched inverse weighted estimator may not always have a smaller variance than the simple inverse weighted estimator). This enriched inverse weighted estimator is not as efficient as the locally efficient estimator and is not doubly robust. The asymptotic variance of this estimator can be estimated by $E_n[\widehat{IF}_t^{EIW}(O;\hat{\mu}_t^{EIW},\hat{\alpha})^2]$ where $\widehat{IF}_t^{EIW}(O;\hat{\mu}_t^{EIW},\hat{\alpha})$ is the same as $IF_t^{EIW}(O;\hat{\mu}_t^{EIW},\hat{\alpha})$ with expectations replaced by empirical averages.

4. Simulation Studies

Sections 4.1 ad 4.2 present simulation studies for binary and continuous outcomes, respectively.

4.1. Binary Outcomes

We assess the finite sample performance of the four proposed estimation procedures for μ_1. In Section 4.1.1, we evaluate the performance when we correctly specify models for the outcome regression and propensity score. In Section 4.1.2, we evaluate the performance under various types of models mis-specification. In our simulations, we only implemented, due to computational complexity, the first two iterations in the computation of the locally efficient estimator (i.e., $j^*=2$). We present simulation results for two sample sizes: 500 and 1000. For each simulation, 2000 datasets were generated and summarized.

4.1.1. Efficiency

In this simulation, S and W were assumed to be independent $Bernoulli(1/2)$ and $Normal(0,1/3)$, respectively. Further, Y_1 and Y_0 given S and W were assumed to be independent Bernoulli distributions with probabilities of success equal to $\text{expit}(1+S+W)$ and $\text{expit}(S+W)$, respectively, and T given X,Y_1 and Y_0 was assumed to be $Bernoulli(\pi_1^*(X))$, where $\pi_1^*(X)=\text{expit}(2S+W+SW)$. Finally, V given T,X,Y_1,Y_0 was assumed to be

$Bernoulli(q_T(S,Y))$, where $q_T(S,Y) = 0.2 + 0.1S + 0.1T + 0.2Y$. Based on these distributions, the true value of μ_1 is 0.8. In the first 8 rows of Table 2, we summarize the results of the simulation. The table show that for each sample size, all estimators are unbiased (see column 3). The Monte Carlo variance (column 4) agrees very well with the average of the estimated variances (column 5) for sample size 1000. For this sample size, the coverage of the estimated 95% confidence intervals are close to their nominal level. For sample size 500, the variance estimator appears to be slightly biased low and the estimated confidence intervals tend to slightly undercover. In terms of efficiency, the simple inverse weighted estimator is very inefficient. The other three estimators, however, perform comparably. It is somewhat surprising to observe that the enriched inverse weighted estimator, which does not require modeling of $\mu_1^*(X)$, performs as well as the locally efficient estimator.

4.1.2. Robustness

We generated data as in the previous subsection, however analyzed them with various mid-specified models. The robust variance was used for the locally efficient estimator and the doubly robust variance was used for the doubly robust estimator. In second set of rows in Table 2, we consider the case where the model for $\mu_t^*(X)$ is misspecified. In our analysis, we incorrectly assumed that $\mu_1(X;\eta) = \text{expit}(\eta_0 + \eta_1 W)$. In this situation, all estimators, as expected, remained unbiased (see column 3). In terms of efficiency, the inverse weighted estimator is the most inefficient and the other three estimators are very comparable (column 4). The coverage rates of the estimated 95% confidence intervals are excellent at both sample sizes (column 6). In the third set of rows in Table 2, we consider the case whether the model for $\pi^*(1,X)$ is misspecified. In our analysis, we incorrectly assumed that $\pi_1(X;\alpha) = \text{expit}(\alpha_0 + \alpha_1 \exp(W))$. As expected, we see that only the doubly robust estimator is unbiased and only its associated confidence interval procedure achieves the nominal coverage rate.

4.2. Continuous Outcomes

We assess the finite sample performance of three proposed estimation procedures for μ_1. We did not implement the locally efficient estimator. In Section 4.2.1, we evaluate the performance when we correctly specify models for the outcome regression and propensity score. In Section 4.2.2, we evaluate the performance under various types of models misspecification. As with the above simulation study, we present simulation results for two sample sizes: 500 and 1000, and, for each simulation, we generated 2000 datasets.

4.2.1. Efficiency

We generated S and W as in Section 4.1.1. We assumed that Y_1 and Y_0, given S and W, followed independent normal distributions with means equal to $0.2 + 0.2S - 0.1W$ and $0.2V - 0.1W$ respectively, and common variance of $1/25$. We assumed that T given X, Y_1, Y_0 was assumed to be $Bernoulli(\pi_1^*(X))$, where $\pi_1^*(X) = expit(2S + W + S \times W)$ and V given T, X, Y_1, Y_0 was assumed to be $Bernoulii(q_T(S,Y))$, where $q_T(S,Y) = 0.2 + 0.1S + 0.1T + 0.2I(Y \geq 0.2)$. Under these distributional assumtions, the true value of μ_1^* is 0.3. The conditional expectations of the form of $E[u(X)|S,Y_t]$ were estimated by fitting linear regressions $u(X) \sim S + Y + S \times Y$ with weights equal to $\frac{VI_{\{T=t\}}}{q_T(S,Y)\pi_t(X;\alpha)}$.

Table 2. Finite-sample properties, efficiency and robustness comparisons for a binary outcome

Sample size	Estimator	$\hat{\mu_1}$	$var \times 10^4$	$\hat{var} \times 10^4$	95% CP(%)
500	SIW	0.80	16.34	15.56	92.60
	EIW	0.80	6.32	6.11	93.90
	DR	0.80	6.38	6.15	94.05
	LE	0.80	6.40	6.11	93.95
1000	SIW	0.80	7.98	7.89	94.35
	EIW	0.80	2.92	3.01	94.50
	DR	0.80	2.94	3.01	94.55
	LE	0.80	2.95	3.01	94.45
$\mu_1^*(X)$ is misspecified					
500	SIW	0.80	16.18	15.57	93.30
	EIW	0.80	6.25	6.14	94.40
	DR	0.80	6.38	6.41	94.90
	LE	0.80	6.38	6.16	94.35
1000	SIW	0.80	7.96	7.91	94.15
	EIW	0.80	3.01	3.03	95.05
	DR	0.80	3.10	3.13	94.95
	LE	0.80	3.09	3.05	94.75
$\pi_1^*(X)$ is misspecified					
500	SIW	0.82	12.23	12.06	87.40
	EIW	0.82	4.69	4.63	80.80
	DR	0.80	6.61	6.60	93.35
	LE	0.81	9.75	4.67	80.80
1000	SIW	0.82	6.14	6.04	83.05
	EIW	0.82	2.36	2.30	70.95
	DR	0.80	3.33	3.28	94.45
	LE	0.81	4.85	2.32	79.70

In the first set of rows of Table 3, we present the results when there is no model mis-specification. In the third column, we see that there is no bias for any of the estimators. Again, we see that the simple inverse weighted estimator is highly inefficient. The doubly robust estimator is the most efficient, but shows only slight improvement over the enriched inverse weighted estimator. The coverage rates of the estimated 95% confidence intervals for all estimation procedures is close to their nominal level. The variance estimator appears biased slighly high.

4.2.2. Robustness

In the second set of rows of Table 3, we consider the effects of mis-specification of the model for $\mu_1^*(X)$. In our analysis, we incorrectly assumed that $\mu_1(X, \eta) = \eta_0 + \eta_1 W$. Here, we see, as expected, that all estimators are unbiased. The enriched inverse weighted estimator is more efficient than the doubly robust estimator in this setting. The simple inverse weighted estimator is the grossly inefficient. The coverage rates for the confidence intervals are excellent, although their are some slight discrepancies between the Monte Carlo variance and the estimated variances.

Finally, in the third set of rows of Table 3, we consider the effects of mis-specification of the model for $\pi_1^*(X)$. In our analysis, we incorrectly assumed that $\pi_1(X, \alpha) = \text{expit}(\alpha_0 + \alpha_1 W)$. Here, only the doubly robust estimator is unbiased and only for this estimator is the coverage rate for the 95% confidence interval near the nominal level.

5. Data Analysis

The NSCOT study was conducted in 15 regions defined by one or more contiguous Metropolitan Statistical Areas (MSAs) located in 12 states. The MSA's were selected from among the 25 largest MSAs located in 19 states for which routinely collected hospital discharge data were available in 1999. MSA's in which the larger non-trauma centers collectively treated fewer than 75 major trauma patients per year as defined by an ICD-derived Injury Severity Score (ICD/ISS) > 15 were excluded.

Within each MSA a sample of level I trauma centers and large non-trauma center hospitals was identified for study inclusion. Hospitals were identified as level I trauma centers if (as of July, 2001) they were designated by a state or regional authority or verified by the ACS/COT. Non-trauma center hospitals were hospitals that were neither designated nor verified as a trauma center at any level and treated at least 25 major trauma patients per year. A total of 27 trauma centers and 124 non-trauma centers were selected. Eighteen (67%) of the trauma centers and 51 (41%) of the non-trauma centers agreed to participate and received approval for the study from their institutional review boards (IRBs).

Eligible for inclusion in the NSCOT study were all trauma patients ages 18-84 treated at one of the participating hospitals for a moderately severe to severe injury (as defined by at least one injury of an Abbreviated Injury Scale (AIS) score of 3 or greater). Excluded were patients who presented with no vital signs and pronounced dead within 30 minutes of arrival at the hospital; patients who did not seek treatment at a hospital within 24 hours of injury; patients 65 years older with a first listed diagnosis of a hip fracture; major burns; patients who were either non-English or non-Spanish speaking; non-U.S. residents; and individuals incarcerated or homeless at the time of injury. Patients recruited into the study were hospitalized over an 18 month period: July, 2001 to November, 2002.

Table 3. Finite-sample properties, efficiency and robustness comparisons for a continuous outcome

Sample size	Estimator	$\hat{\mu_1}$	$var \times 10^4$	$\hat{var} \times 10^4$	95% CP(%)
500	SIW	0.30	4.10	4.14	94.25
	EIW	0.30	1.79	2.02	95.75
	DR	0.30	1.72	1.86	95.85
1000	SIW	0.30	1.98	2.05	94.50
	EIW	0.30	0.95	1.02	96.35
	DR	0.30	0.92	0.99	96.45
$\mu_1^*(X)$ is misspecified					
500	SIW	0.30	4.15	4.17	94.45
	EIW	0.30	2.00	2.17	96.55
	DR	0.30	2.52	2.54	94.75
1000	SIW	0.30	1.98	2.07	95.20
	EIW	0.30	0.76	0.84	95.90
	DR	0.30	1.03	0.98	94.45
$\pi_1^*(X)$ is misspecified					
500	SIW	0.33	3.50	3.69	70.20
	EIW	0.33	1.99	2.17	40.55
	DR	0.30	1.81	1.78	93.90
1000	SIW	0.33	1.80	1.84	46.80
	EIW	0.33	0.73	0.79	13.25
	DR	0.30	0.77	0.74	94.20

The patient sample was selected and eligibility determined in two stages, First, administrative discharge records and emergency department (ED) logs were prospectively reviewed to identify all patients ages 18-84 who either died in the ED or were discharged alive or dead from the hospital with a certain principal ICD9 CM diagnosis codes. A computerized mapping of ICD-9 CM discharge diagnoses to AIS severity scores was then applied to select patients with at least one diagnosis corresponding to an ICD/AIS score of 3 or greater. A total of 18,198 patients across the 69 hospitals met these initial eligibility criteria.

In the second stage of the sampling process, all 1,438 hospital deaths and a random sample of 8,021 patients discharged alive, stratified within hospitals on (i) age: 18-64 and 65 to 84; (ii) ICD/ISS severity scores: $<= 15$ and > 15; and (iii) principal body region injured: head injury of ICD/AIS 3-6 regardless of other injuries; no head injury of ICD/AIS 3-6 but one or more extremity injures of ICD/AIS 3-6; all others with at least one injury of ICD/AIS 3-6. Stratifying the sample of live hospital discharges was necessary to facilitate balance in baseline risk between trauma centers and non-trauma centers.

Medical records were obtained for 1,391 (97%) of the hospital deaths. On the basis of a detailed medical record review, 287 of these deaths were further excluded because they failed to meet the inclusion criteria, leaving 1,104 deaths who were eligible and for whom complete medical record data were abstracted.

Patients discharged alive and selected for the study were contacted by mail at three months post-injury. Those who did not refuse by mail were contacted by phone and consent obtained for access to the medical record and interviews at three and twelve months. Of the 8,021 live discharges selected for the study, 4,866 (61%) were enrolled (1,635 could not be located; 1,177 refused to participate; and 343 completed the three month interview but never provided written permission to access their medical record). Of those enrolled, 779 were determined ineligible upon further review of their medical record (usually because their discharge diagnoses did not meet eligibility criteria), leaving 4,087 live discharges who were eligible and for whom complete medical record data were abstracted.

In our analysis, Y denotes of the indicator of in-hospital death and S denotes the hospital/age/severity/body-region strata. Note that T is implicitly encoded in S. We assume that, for initially eligible individuals, the abstraction process is unrelated to the true eligibility of the patient. We also assume that if an individual is not initially eligible then the individual is not truly eligible. Under these assumptions, the probability that a truly eligible individual is included in the dataset is equal to the inverse of the Y/S-specific proportion of medical records abstracted. On patients included in the dataset, we have medical record covariates (W), which will be used to control for selection bias due to non-random trauma center assignment.

There are two differences in the sampling design of NSCOT and the one used in this paper: (1) the sampling strata is defined at individual hospital level, not treatment level (e.g. trauma/non-trauma centers) (2) individuals in the validation sample may not be truly eligible. The simple inverse weighted estimator can directly handle these two differences, while the other three estimators cannot. In our analysis, we aggregate hospital-level strata into trauma/non-trauma strata and re-calculate the sampling weights. The locally efficient, doubly robust and enriched inverse weighted estimators can be extended to handle the truly eligibility problem if medical record covariates (W) were available for subjects in the validation sample who were not truly eligible. However, this information was not routinely collected in the NSCOT study. For purpose of illustration, we address the eligibility problem, by estimating, within strata, the number of truly eligible individuals among those not validated (using the assumptions described in the previous paragraph). We then consider

Table 4. In-Hospital Mortality from Different Estimators

	Observed	SIW	EIW	DR	LE
Trauma center(%)	7.88	7.64	7.55	7.47	7.43
		(0.28)	(0.25)	(0.25)	(0.25)
Non-trauma center(%)	5.77	11.18	10.82	10.19	10.53
		(2.36)	(2.36)	(1.81)	(2.09)
Relative risk	1.37	0.68	0.70	0.73	0.71
		(0.40-0.97)	(0.40-1.00)	(0.48 - 0.99)	(0.43-0.98)

the estimated number of individuals as part of the non-validation sample. At the end of this process, a total number of $n = 15,617$ individuals were considered as part of our dataset. Our analysis does not consider n to be a random variable. Thus, our estimators may be biased and we will likely be under-reporting the true level of uncertainty.

We used the same propensity score model as in MacKenzie et al. (2006). Separate regression models were fit for $E[Y_1|X]$ and $E[Y_0|X]$. The models included main effect terms for the following covariates: age (natural cubic spline), gender, race, Charlson score, mechanism of injury, first ED shock, field GCS motor, first ED GCS motor,maximum AIS, maximum AIS for head injury, thorax injury, abdomen injury and extremity injury, first ED Pupils, coagulopathy, obesity, flail chest, open skull, paralysis, long bone fx or amputation, midline shift and insurance status. An interaction between mechanism of injury and insurance status was also included.

The estimated in-hospital mortality and relative risk using four different estimators are shown in Table 4. The four different estimators of in-hospital mortality under trauma care are similar and comparable to the observed mortality among those treated at trauma centers. The standard errors are also comparable. For non-trauma center care, the estimators of in-hospital mortality are almost twice as large as the observed mortality among those treated at non-trauma centers. The doubly robust estimator yields the lowest of the point estimates and also have the lowest standard error. All four estimation procedures yield similar relative risk estimates, with the doubly robust estimator providing the shortest confidence interval.

6. Discussion

In this paper, we presented four estimators for the causal effect of a binary treatment in the two-phase outcome-dependent sampling design. Based on simulation and empirical work, we recommend use of the doubly robust estimator because of its reasonable efficiency, robustness and its computational tractability. The enriched inverse weighted estimator also performed well and we recommend its use as well. The locally efficient estimator is very computationally intensive and we have found no evidence that it performs better than the doubly robust or enriched inverse weighted estimators. The simple inverse weighted is inefficient, but it does have the advanatge of being able to handle sampling complexities that we could not handle easily with our other estimators.

There are several directions for future research. First, in this paper, we assumed that the sampling probability $q_T(S, Y)$ is known. However, in many studies the sampling weights are not known apriori and must be estimated. Even if they are known, it can be shown that treating them as estimates can yield more efficent estimators. In doing so, the influence functions no longer take the form of Equation (1). To derive the new class of influence functions, one takes the infleunce functions in Equation (1) and subtracts their projection

onto the space spanned by the score function from a model for the sampling probability. It is important to note, however, that the score function is not well defined when the the sampling probability is one in some strata, which happens often in outcome-dependent sampling designs.

Second, in presence of model mis-specification, the EIW, DR and LE estimators may not have smaller variances than the SIW estimator. Tan (2006) introduced a way to get a doubly roubst estimator which always improves over SIW. Future work could extend his method to outcome-dependent sampling designs.

In this paper, we have used the i.i.d. framework for semi-parametric inference developed by Bickel et al. (1993). The NSCOT study does not fit ideally into this framework. This stems from the fact that NSCOT employed Binomial sampling within strata. Future work should focus on causal inference methods under Binomial and other sampling schemes.

7. Appendix

Proposition 7.1. *Define* $\epsilon_1 = Y_1 - \mu_1^*$, $\epsilon_0 = Y_0 - \mu_0^*$ *and reparameterize the full data* (X, Y_0, Y_1) *to* $(X, \epsilon_0, \epsilon_1)$. *The nuisance tangent space of the full data is*

$$\begin{aligned}\Lambda^F = \{a(\epsilon_0, \epsilon_1, X): \quad & E[a(\epsilon_0, \epsilon_1, X)] = 0, \\ & E[\epsilon_0 a(\epsilon_0, \epsilon_1, X)] = 0, \\ & E[\epsilon_1 a(\epsilon_0, \epsilon_1, X)] = 0\}\end{aligned} \tag{8}$$

The orthogonal nuisance tangent space of the full data is

$$\Lambda^{F,\perp} = \{B^{2\times 2}\begin{pmatrix}\epsilon_0 \\ \epsilon_1\end{pmatrix} : B \text{ is any } 2\times 2 \text{ matrix of constants}\} \tag{9}$$

Proof: The joint distribution $f(X, \epsilon_1, \epsilon_0)$ is any distribution satisfying $E[\epsilon_1] = 0$ and $E[\epsilon_0] = 0$. The result follows. □

Under the assumption of *no unmeasured confounders* and that the sampling probability only depends on (S, T, Y) , the complete data $L = (X, Y_0, Y_1, V, T)$ has the likelihood

$$f(X, Y_0, Y_1)P(T|X, Y_0, Y_1)P(V|T, X, Y_0, Y_1) = f(X, Y_0, Y_1)P(T|X)P(V|T, S, Y).$$

Thus, the nuisance tangent space of the complete data likelihood can be factorized as

$$\Lambda^{comp} = \Lambda^F \oplus \Lambda^{T|X} \oplus \Lambda^{V|T,S,Y}.$$

Since $P(V|T, S, Y)$ is known by study design, we have

$$\Lambda^{comp} = \Lambda^F \oplus \Lambda^{T|X}.$$

We will also assume that the propensity score $\pi_t(X)$ is known up to a finite number of parameters with α^* denoting the true parameter, i.e.

$$\pi_t^*(X) = \pi_t(X; \alpha^*).$$

Without loss of generality, we will assume that

$$\text{logit } \pi_1(X; \alpha^*) = l(X, \alpha^*)$$

where $l(X, \alpha^*)$ is a specified function of X and α which is differentiable in α. This model implies that $\pi_t(X; \alpha^*) = \exp(t \cdot l(X; \alpha^*))/(1 + \exp(l(X; \alpha^*)))$.

Table 5. Missing patterns

(V,T)	O	$P(V,T\|X,Y_0,Y_1)$
$(0,1)$	(S,Y_1)	$(1-q_1(S,Y_1))\pi_1^*(X)$
$(0,0)$	(S,Y_0)	$(1-q_0(S,Y_0))\pi_0^*(X)$
$(1,1)$	(X,Y_1)	$q_1(S,Y_1)\pi_1^*(X)$
$(1,0)$	(X,Y_0)	$q_0(S,Y_0)\pi_0^*(X)$

Proposition 7.2.

$$\Lambda^{T|X} = \{B^{2\times r}\frac{\partial l(X;\alpha^*)}{\partial \alpha}(T-\pi_1(X;\alpha^*)) : \textit{for all } B\}.$$

Proof: The likelihood is

$$p_{T|X}(t,x;\alpha) = \pi_1(x;\alpha)^t(1-\pi_1(x;\alpha))^{1-t}$$

The log-likelihood is equal to

$$\log p_{T|X}(t,x;\alpha) = t\log\pi_1(x;\alpha) + (1-t)\log(1-\pi_1(x;\alpha))$$

The score function is

$$S_\alpha = \left(\frac{t}{\pi_1(x;\alpha)} - \frac{1-t}{1-\pi_1(x;\alpha)}\right)\frac{\partial\pi_1(x;\alpha^*)}{\partial\alpha} = \frac{\partial l(x;\alpha^*)}{\partial\alpha}(t-\pi_1(x;\alpha)).$$

Since $\Lambda^{T|X}$ is the space spanned by S_α, the result follows. □

The observed data is $O = (S, VW, TY_1, (1-T)Y_0, V, T)$, indicating that there is missing data. The missing patterns are listed in Table 5. The data are not missing at random because the probability of missingness depends on W which is not always observed.

Let $\Lambda_1 = \Lambda^F$, $\Lambda_2 = \Lambda^{T|X}$. The nuisance tangent space for the observed data is equal to

$$\Lambda^O = \Lambda_1^O + \Lambda_2^O,$$

where $\Lambda_j^O = \overline{Range(g\circ\Pi_j)}, j = 1,2$, $Range(\cdot)$ is the range of an operator.$g : \Omega^L \to \Omega^O$, $g(\cdot) = E[\cdot|O]$, Ω^L and Ω^O are spaces of mean 0 random functions of L and O, respectively, Π_j is the projection operator from Ω^L to Λ_j and $\bar{S}$ denote the closed linear span of the set S (Bickel et al., 1993). Since the data are not missing at random, Λ_1^O and Λ_2^O are not orthogonal. Rgardless, we still have the following results:

$$\Lambda^{O,\perp} = \Lambda_1^{O,\perp} \cap \Lambda_2^{O,\perp}$$

Rotnitzky and Robins (1997) showed how to compute $\Lambda_1^{O,\perp}$:

Proposition 7.3.

$$\Lambda_1^{O,\perp} = \left\{B^{2\times 2}\begin{pmatrix} \frac{VT}{q_1(S,Y_1)\pi_1(X;\alpha^*)}(Y_1-\mu_1^*) \\ \frac{V(1-T)}{q_0(S,Y_0)\pi_0(X;\alpha^*)}(Y_0-\mu_0^*)\end{pmatrix} + A^{(2)} :\ \textit{for all } B, A^{(2)}\in\Lambda^{(2)}\right\},$$

where $\Lambda^{(2)} = \{b(O) : E[b(O)|L] = 0\}$.

Proposition 7.4.

$$\Lambda^{(2)} = \left\{ \sum_{\tau=0}^{1} \phi_{2,\tau}(O; g_\tau) + \phi_3(O; \alpha^*, h) : \textit{ for all } g_o, g_1, h \right\}$$

where

$$\begin{aligned} \phi_{2,\tau}(O; g_\tau) &= 1_{\{T=\tau\}} \left(1 - \frac{V}{q_\tau(S, Y_\tau)}\right) g_\tau(S, Y_\tau) \\ \phi_3(O; \alpha^*, h) &= \frac{V}{q_T(S, Y_T)} \left(\frac{T - \pi_1(X; \alpha^*)}{R(X; \alpha^*)}\right) h(X), \end{aligned}$$

g_1, g_0, h are arbitrary 2×1 vectors of functions of their arguments and $R(X; \alpha^*) = \pi_1(X; \alpha^*)\pi_0(X; \alpha^*)$.

Proof: Consider a function of the observed data

$$\begin{aligned} b(O) =& (1 - V)Tg_1(S, Y_1) \\ &+ (1 - V)(1 - T)g_2(S, Y_0) \\ &+ VTg_3(X, Y_1) \\ &+ V(1 - T)g_4(X, Y_0) \end{aligned}$$

Then,

$$\begin{aligned} E[b(O)|L] =& (1 - q_1(S, Y_1))\pi_1^*(X)g_1(S, Y_1) \\ &+ (1 - q_0(S, Y_0))\pi_0^*(X)g_2(S, Y_0) \\ &+ q_1(S, Y_1)\pi_1^*(X)g_3(X, Y_1) \\ &+ q_0(S, Y_0)\pi_0^*(X)g_4(X, Y_0) \end{aligned}$$

Seting the expectation to zero, we have

$$\begin{aligned} &(1 - q_1(S, Y_1))\pi_1^*(X)g_1(S, Y_1) + q_1(S, Y_1)\pi_1^*(X)g_3(X, Y_1) \\ =& - (1 - q_0(S, Y_0))\pi_0^*(X)g_2(S, Y_0) - q_0(S, Y_0)\pi_0^*(X)g_4(X, Y_0) \end{aligned}$$

The left hand side is a function of (X, Y_1) and the right hand side is a function of (X, Y_0). For them to be equal, both have to be a function of X alone. Thus,

$$(1 - q_1(S, Y_1))\pi_1^*(X)g_1(S, Y_1) + q_1(S, Y_1)\pi_1^*(X)g_3(X, Y_1) = h(X)$$

$$-(1 - q_0(S, Y_0))\pi_0^*(X)g_2(S, Y_0) - q_0(S, Y_0)\pi_0^*(X)g_4(X, Y_0) = h(X)$$

where $h(X)$ is a function of X. Then,

$$g_3(X, Y_1) = -\frac{1 - q_1(S, Y_1)}{q_1(S, Y_1)} g_1(S, Y_1) + \frac{1}{q_1(S, Y_1)} \frac{h(X)}{\pi_1^*(X)}$$

and

$$g_4(X, Y_0) = -\frac{1 - q_0(S, Y_0)}{q_0(S, Y_0)} g_2(S, Y_0) - \frac{1}{q_0(S, Y_0)} \frac{h(X)}{\pi_0^*(X)}$$

Simple algebra leads to the proposition. □

Scharfstein et al. (1999) showed the following result:

Proposition 7.5.

$$\begin{aligned}\Lambda_2^{O,\perp} &= \{b(O) : \Pi[b(O)|\Lambda^{T|X}] = 0\} \\ &= \{b(O) : b(O) \in \Lambda^{T|X,\perp}\}\end{aligned}$$

For the simplicity of notation, we consider the estimation of μ_t from now on.

Proposition 7.6. *The orthogonal nuisance tangent space of the observed data is equal to*

$$\Lambda^{O,\perp} = \{B\phi_1(O;\mu_t^*,\alpha^*,\mu_t^*(X)) + \sum_{\tau=0}^{1}\phi_{2,\tau}(O;g_\tau) + \phi_3(O;\alpha^*,b),: \text{ for all } B\}$$

where

$$\begin{aligned}\phi_1(O;\mu_t^*,\alpha^*,\mu_t^*(X)) =& \frac{VI_{\{T=t\}}}{q_t(S,Y_t)\pi_t(X;\alpha^*)}(Y_t-\mu_t^*) \\ &- (-1)^t\frac{V}{q_T(S,Y_T)}\left(\frac{T-\pi_1(X;\alpha^*)}{R(X;\alpha^*)}\right)\pi_{1-t}(X;\alpha^*)(\mu_t^*(X)-\mu_t^*) \\ \phi_{2,\tau}(O;g_\tau) =& 1_{\{T=\tau\}}\left(1-\frac{V}{q_\tau(S,Y_\tau)}\right)g_\tau(S,Y_\tau) \\ \phi_3(O;\alpha^*,b) =& \frac{V}{q_T(S,Y_T)}\left(\frac{T-\pi_1(X;\alpha^*)}{R(X;\alpha^*)}\right)b(X),\end{aligned}$$

g_1, g_0 are arbitrary functions of their arguments, $b(X) \in \mathcal{T}^\perp$ and

$$\mathcal{T} = \left\{k'\frac{\partial l(X;\alpha^*)}{\partial\alpha} : k \text{ is an arbitrary constant vector}\right\}$$

Proof: We take an element from $\Lambda_1^{O,\perp}$ and find out what conditions it must satisfy so that it also belongs to $\Lambda_2^{O,\perp}$. Note that $\phi_{2,\tau}(O;g_\tau) \in \Lambda_2^{O,\perp}$ and that for any $b(O) \in \Lambda_1^{O,\perp}$, we have

$$\begin{aligned}& E\left[b(O)\frac{\partial l(X;\alpha^*)}{\partial\alpha}(T-\pi_1(X;\alpha^*))\right] \\ =& E\left[\left(\frac{VI_{\{T=t\}}}{q_t(S,Y_t)\pi_t(X;\alpha^*)}(Y_t-\mu_t^*)+\phi_3(O;\alpha^*,h_t)\right)\frac{\partial l(X;\alpha^*)}{\partial\alpha}(T-\pi_1(X;\alpha^*))\right] \\ =& E\left[\frac{\partial l(X;\alpha^*)}{\partial\alpha}\{(-1)^{t-1}(\mu_t^*(X)-\mu_t^*)\pi(1-t,X)+h_t(X)\}\right]\end{aligned}$$

Therefore,

$$h_t(X) = (-1)^t\pi_{1-t}(X;\alpha^*)(\mu_t^*(X)-\mu_t^*) + b(X),$$

where $E[\frac{\partial l(X;\alpha^*)}{\partial\alpha}b(X)] = 0$, i.e., $b(X) \in \mathcal{T}^\perp$. □

Proposition 7.7. *The set of influence functions of all RAL estimator of μ_t^* is*

$$\Lambda_*^{O,\perp} = \{\phi_1(O;\mu_t^*,\alpha^*,\mu_t^*(X)) + \sum_{\tau=0}^{1}\phi_{2,\tau}(O;g_\tau) + \phi_3(O;\alpha^*,b)\}$$

Proof: The set of influence functions of all RAL estimator is equal to

$$\Lambda_*^{O,\perp} = \{g(O;\mu_t^*,\alpha^*,\mu_t^*(X)) \in \Lambda^{O,\perp} : E[\frac{\partial gO;\mu_t^*,\alpha^*,\mu_t^*(X))}{\partial \mu_t^*}] = -1\}$$

It is straightforward to show that for any $gO;\mu_t^*,\alpha^*,\mu_t^*(X)) \in \Lambda^{O,\perp}$,

$$E[\frac{\partial gO;\mu_t^*,\alpha^*,\mu_t^*(X))}{\partial \mu_t^*}] = -B.$$

Therefore, B should be 1. □

Proposition 7.8. *For an arbitrary function of observed data $h(O)$,*

$$\Pi[h(O)|\Lambda_g] = \sum_{\tau=0}^{1} I_{\{T=\tau\}} \left(1 - \frac{V}{q_\tau(S,Y_\tau)}\right) \frac{E\left[h(O)I_{\{T=\tau\}}\left(1-\frac{V}{q_\tau(S,Y_\tau)}\right) \,\middle|\, S,Y_\tau\right]}{E\left[I_{\{T=\tau\}}\left(1-\frac{V}{q_\tau(S,Y_\tau)}\right)^2 \,\middle|\, S,Y_\tau\right]}$$

Proof: Suppose $\Pi[h(O)|\Lambda_{g_\tau}] = \phi_{2,\tau}(O;g_\tau^*)$, then for any g_τ,

$$E[(h(O) - \phi_{2,\tau}(O;g_\tau^*))\phi_{2,\tau}(O;g_\tau)] = 0.$$

g_τ^* can be directly solved from the above equation. □

Proposition 7.9. *For an arbitrary function of observed data $h(O)$,*

$$\Pi[h(O)|\Lambda_b] = \frac{V}{q_T}\left(\frac{T-\pi_1(X;\alpha^*)}{R(X;\alpha^*)}\right) A(X)^{-1}\left\{C_h(X) - k_h' \frac{\partial l(X;\alpha^*)}{\partial \alpha}\right\}, \tag{10}$$

where

$$C_h(X) = E\left[h(O)\frac{V}{q_T(X,Y_T)}\frac{T-\pi_1(X;\alpha^*)}{R(X;\alpha^*)}\,\middle|\, X\right],$$

$$A(X) = \pi_1(X;\alpha^*)^{-1}E\left[q_1(S,Y_1)^{-1}\,|\,X\right] + \pi_0(X;\alpha^*)^{-1}E\left[q_0(S,Y_0)^{-1}\,|\,X\right]$$

and

$$k_h' = E\left[A(X)^{-1}C_h(X)\frac{\partial l(X;\alpha^*)}{\partial \alpha}'\right] E\left[A(X)^{-1}\frac{\partial l(X;\alpha^*)}{\partial \alpha}\frac{\partial l(X;\alpha^*)}{\partial \alpha}'\right]^{-1}$$

Proof: Suppose $\Pi[h(O)|\Lambda_b] = \phi_3(O;\alpha^*,b^*)$, b^* can be determined by the following two restrictions: (1) for any b, $E[(h(O) - \phi_3(O;\alpha^*,b^*))\phi_3(O;\alpha^*,b)] = 0$ (2) $b^* \in \mathcal{T}^\perp$. □

References

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semi Parametric Models.* The Johns Hopkins University Press, Baltimore, Maryland.

Breslow, N., McNeney, B., and Wellner, J. A. (2003). Large sample theory for semiparametric regression models with two-phase, outcome dependent sampling. *The Annals of Statistics* **31,** 1110–1139.

Breslow, N. E. (2000). Statistics in the life and medical sciences. *Journal of the American Statistical Association* **95,** 281–282.

Breslow, N. E. and Cain, K. C. (1988). Logistic regression for two-stage case-control data. *Biometrika* **75,** 11–20.

Breslow, N. E. and Holubkov, R. (1997a). Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. *Journal of the Royal Statistical Society, Series B: Methodological* **59,** 447–461.

Breslow, N. E. and Holubkov, R. (1997b). Weighted likelihood, pseudo-likelihood and maximum likelihood methods for logistic regression analysis of two-stage data. *Statistics in Medicine* **16,** 103–116.

Carroll, R. J. and Wand, M. (1991). Semiparametric estimation in logistic measurement error models. *J. R. Statist. Soc. B* **53,** 573–585.

Chatterjee, N., Chen, Y.-H., and Breslow, N. E. (2003). A pseudoscore estimator for regression problems with two-phase sampling. *Journal of the American Statistical Association* **98,** 158–168.

Cochran, W. G. (1963). *Sampling Techniques*, page 412. Wiley.

Cosslett, S. R. (1981). Maximum likelihood estimator for choice-based samples. *Econometrica* **49,** 1289–1316.

Cosslett, S. R. (1983). Distribution-free maximum likelihood estimator of the binary choice model. *Econometrica* **51,** 765–782.

Fears, T. R. and Brown, C. C. (1986). Logistic regression methods for retrospective case-control studies using complex sampling procedures. *Biometrics* **42,** 955–960.

Lawless, J. F., Kalbfleisch, J. D., and Wild, C. J. (1999). Semiparametric methods for response-selective and missing data problems in regression. *Journal of the Royal Statistical Society, Series B: Statistical Methodology* **61,** 413–438.

MacKenzie, E. J., Rivara, F. P., Jurkovich, G. J., Nathens, A. B., Frey, K. P., Egleston, B. L., Salkever, D. S., and Scharfstein, D. O. (2006). A national evaluation of the effect of trauma-center care on mortality. *The new england journal of medicine* **354,** 366–378.

Newey, W. K. (1994). The asymptotic variance of semiparametric estimators. *Econometrica* **62,** 1349–1382.

Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association* **33,** 101–116.

Pepe, M. S. and Fleming, T. R. (1991). A nonparametric method for dealing with mismeasured covariate data. *Journal of the American Statistical Association* **86,** 108–113.

Robins, J. M., Hsieh, F., and Newey, W. (1995). Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates. *Journal of the Royal Statistical Society, Series B (Methodological)* **57,** 409–424.

Robins, J. M. and Ritov, Y. (1997). Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. *Statistics in Medicine* **16,** 285–319.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89,** 846–866.

Rotnitzky, A. and Robins, J. M. (1997). Analysis of semiparametric regression models with non-ignorable non-response. *Statistics in Medicine,* **16,** 81–102.

Rubin, D. B. (1986). Comments on "Statistics and causal inference". *Journal of the American Statistical Association* **81,** 961–962.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Reply to comments on "Adjusting for nonignorable drop-out using semiparametric nonresponse models". *Journal of the American Statistical Association* **94,** 1135–1146.

Schill, W., Jockel, K.-H., Joeckel, K.-H., Jöckel, K.-H., Drescher, K., and Timm, J. (1993). Logistic analysis in case-control studies under validation sampling. *Biometrika* **80,** 339–352.

Scott, A. J. and Wild, C. J. (1991). Fitting logistic regression models in stratified case-control studies. *Biometrics* **47,** 497–510.

Scott, A. J. and Wild, C. J. (1997). Fitting regression models to case-control data by maximum likelihood. *Biometrika* **84,** 57–71.

Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association* **101,** 1619–1637.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. Springer, New York.

van der Laan, M. J. and Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer Verlag, New York.

Weaver, M. A. and Zhou, H. (2005). An estimated likelihood method for continuous outcome regression models with outcome-dependent sampling. *Journal of the American Statistical Association* **100,** 459–469.

White, J. E. (1982). A two stage design for the study of the relationship between a rare exposure and a rare disease. *American Journal of Epidemiology* **115,** 119–128.